Kubernetes in the CERN Cloud

HEPiX Autumn 2020

Thomas Hartland

New Features

New Kubernetes versions

- Three new versions this year
  - v1.17
  - v1.18
  - v1.19
- New Kubernetes features are nice, but each new version also brings better integration with OpenStack.

Node groups

- Create heterogeneous clusters
  - Different node sizes
  - Have some GPU nodes in a CPU cluster
  - Distribute nodes across availability zones
- Kubernetes nodes are labelled with their node group name.
  - Node selectors to put pods on a specific node group.
  - Or pod topology constraints to spread evenly across groups.

Cluster autoscaling

- On K8s v1.19, the cluster autoscaler supports node groups.
- The autoscaler can scale node groups individually.
- Node group autodiscovery makes it easier to add new node groups to be autoscaled.
  - The min/max size that the autoscaler respects for each node group can be updated through OpenStack.

Loadbalancers

- Using the new SDN features at CERN.
- Now possible to use `type: LoadBalancer` services.
  - Will by default provision a new loadbalancer
  - or can specify an existing loadbalancer to join as a member.
- Alternative to ingress for exposing cluster services.

Multi-master clusters

- SDN was also a prerequisite for this.
  - Kubernetes API is accessed through a loadbalancer.
- Ensures that a cluster's Kubernetes API will be accessible as long as one node is up.
- The etcd data store is also replicated, so data loss is much less likely.

Multi-master clusters

- We are offering this with a separate template currently

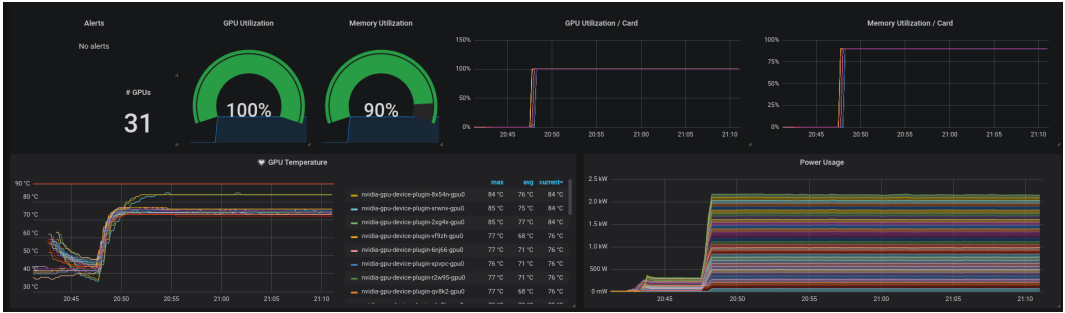
```
$ openstack coe cluster create \
      --cluster-template kubernetes-1.18.6-3-multi \
      --master-count 3 \
      ....
```

- Eventually this will be possible on the standard templates.

GPUs

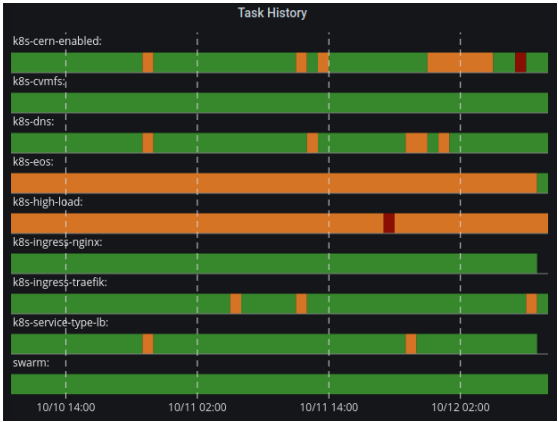
- If you have access to GPU flavors, create a cluster/nodegroup using that flavor.
- GPUs drivers are automatically initialised if the label `nvidia_gpu_enabled=true` is passed at cluster creation.
- In-cluster monitoring dashboards are also created.
- Limited to 1 pod = 1 GPU at the moment, no vGPUs yet.

GPUs


Alerts
No alerts
GPUs
31
GPU Utilization
100%
Memory Utilization
90%
GPU Utilization / Card
150%
100%
50%
0%
20:45
20:50
20:55
21:00
21:05
21:10
Memory Utilization / Card
100%
75%
50%
25%
0%
20:45
20:50
20:55
21:00
21:05
21:10
GPU Temperature
90 °C
80 °C
70 °C
60 °C
50 °C
40 °C
30 °C
20:45
20:50
20:55
21:00
21:05
21:10
max
avg
current
nvidia-gpu-device-plugin-8x54n-gpu0 84 °C 76 °C 84 °C
nvidia-gpu-device-plugin-xrwnv-gpu0 85 °C 75 °C 84 °C
nvidia-gpu-device-plugin-2xp4x-gpu0 85 °C 77 °C 84 °C
nvidia-gpu-device-plugin-vfljzh-gpu0 77 °C 68 °C 76 °C
nvidia-gpu-device-plugin-6nj66-gpu0 77 °C 71 °C 76 °C
nvidia-gpu-device-plugin-xpvpc-gpu0 76 °C 71 °C 76 °C
nvidia-gpu-device-plugin-r2w95-gpu0 77 °C 71 °C 76 °C
nvidia-gpu-device-plugin-qv9k2-gpu0 77 °C 68 °C 76 °C
Power Usage
2.5 kW
2.0 kW
1.5 kW
1.0 kW
500 W
0 mW
20:45
20:50
20:55
21:00
21:05
21:10

Automated testing

- Testing process
  - Create a cluster
  - Run a Job in that cluster.
  - Check the status of the Job.



Automated testing

Percent of tests passing

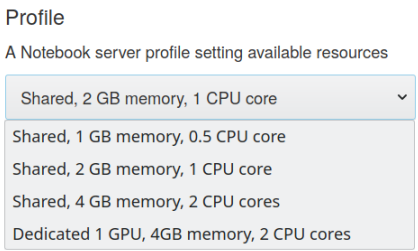
	k8s-cern-enabled	k8s-cvmfs	k8s-dns	k8s-eos	k8s-high-load	k8s-ingress-nginx	k8s-ingress-traefik	k8s-service-type-lb
kubernetes-1.19.0-2	85%	100%	92%	100%	0%	100%	100%	96%
kubernetes-1.18.6-3	92%	100%	100%	96%	2%	100%	100%	100%
kubernetes-1.17.9-2	-	100%	96%	0%	6%	100%	100%	-
kubernetes-1.15.3-3	-	100%	100%	0%	98%	-	94%	-

- We will make these parts of the dashboard available to users.

Use cases at CERN

Jupyter notebooks

- Multiple jupyter notebook services running at CERN.
- Good use for clusters with mixed CPU and GPU nodes.
- Can launch environments for
  - Machine learning
  - Distributed computing
- Or use Binder
  - Generates a notebook container image from a git repository.



Rucio

- Migrating from Puppet managed VMs to Kubernetes.
- Current setup is stable, handling 50% of Rucio load.
  - Moving the daemon component fully to K8s in the next weeks.
- Deployment with helm/flux.
  - Makes it easier for other experiments/organisations to set up Rucio.
- Webinar: https://indico.cern.ch/event/915701/

ATLAS batch processing

- Testing Kubernetes clusters as part of the WLCG.
- Creating grid pilots as Kubernetes jobs.
- Numbers from April:
  - 86 node cluster in CERN cloud
  - 700 cores
- Getting an update in a webinar next week: https://indico.cern.ch/event/950884/

OpenStack control plane

- Puppet deployment for APIs, conductors, rabbitMQ on VMs.
- Containerised deployment:
  - Multi-master, multi-AZ cluster.
  - Multiple external loadbalancers.
  - Using OpenStack helm charts and managed with GitOps.
- First, added the cluster to the existing Magnum loadbalancer.
  - Automated tests found some issues, so we removed it from the LB.
  - We have a test region (fully containerised), running tests here as well.
- Webinar: https://indico.cern.ch/event/915714/

The future

Magnum

- Finalise in place upgrades.
  - Patch releases first, minor releases later.
- Scale nodegroups to 0.
- Support vGPUs.
- Enable cluster auth/authz with CERN OIDC.
- Centralised container vulnerability scanning.

We are looking at

Questions?

Extra slides

Helm / GitOps

- We have been encouraging use of Helm for application deployment, and using it ourselves.
- So far very happy with GitOps workflow using Flux.
  - Automatically apply config changes from repository to cluster.
  - Secret management was difficult to start with but SOPS makes this easier.
  - SOPS integrates with secret store services, we added support for OpenStack Barbican.

Weblogic

- Running ~100 of CERN's administrative services.
- From Puppet VMs to multiple Kubernetes clusters.
- One application per K8s namespace.
  - Easy separation of test/prod environments.
- Time to deploy a new application is massively reduced.
- Migration is still ongoing.
- Webinar: https://indico.cern.ch/event/944703/